Appendix B. Source and Reliability of Data

SOURCE OF DATA

The estimates in this report are based primarily on data obtained in October 1984 from the Current Population Survey (CPS) conducted by the Bureau of the Census and from supplementary questions to the CPS. The monthly CPS deals mainly with labor force data for the civilian noninstitutional population. Questions relating to labor force participation are asked about each member in every sample household. In addition, in October 1984, supplementary questions were asked about computer use. For this report, persons in the Armed Forces living off post or with their families on post are also included.

Current Population Survey (CPS). The present CPS sample was selected from the 1980 decennial census files with coverage in all 50 States and the District of Columbia. The sample is continually updated to reflect new construction. The 1984 CPS sample was located in 629 areas comprising 1,148 counties, independent cities, and minor civil divisions in the Nation. In this sample, approximately 61,500 occupied households were eligible for interview. Of this number, about 2,500 occupied units were visited but interviews were not obtained because the occupants were not found at home after repeated calls or were unavailable for some other reason.

CPS estimation procedure. The estimation procedure used in this survey involved the inflation of the weighted sample results to independent estimates of the total civilian noninstitutional population of the United States by age, race, sex, and Hispanic/ non-Hispanic categories. These independent estimates are based on statistics from the 1980 Decennial Census of Population; statistics on births, deaths, immigration, emigration; and statistics on the strength of the Armed Forces.

RELIABILITY OF ESTIMATES

Since the CPS estimates were based on a sample, they may differ somewhat from the figures that would have been obtained if a complete census had been taken using the same questionnaires, instructions, and enumerators. There are two types of errors possible in an estimate based on a sample survey: sampling and nonsampling. The accuracy of a survey result depends on both types of errors, but the full extent of the nonsampling error is unknown. Consequently, particular care should be exercised in the interpretation of figures based on a relatively small number of cases or on small differences between estimates. The standard errors provided for the CPS estimates primarily indicate the magnitude of the sampling error. They also partially measure the effect of some nonsampling errors in responses and enumeration, but do not measure any systematic biases in the data. (Bias is the difference, averaged over all possible samples, between the sample estimates and the desired value.)

Nonsampling variability. Nonsampling errors can be attributed to many sources, e.g., inability to obtain information about all cases in the sample, definitional difficulties, differences in the interpretation of questions, inability or unwillingness on the part of respondents to provide correct information, inability to recall information, errors made in collection such as in recording or coding the data, errors made in processing the data, errors made in estimating values for missing data, and failure to represent all units with the sample (undercoverage).

Undercoverage in the CPS results from missed housing units and missed persons within sample households. Overall undercoverage, as compared with the level of the 1980 decennial census, is about 7 percent. It is known that CPS undercoverage varies with age, sex, and race. Generally, undercoverage is larger for males than for females and larger for Blacks and other races combined than for Whites. Ratio estimation to independent age-sex-race-Hispanic population controls, as described previously, partially corrects for the bias due to survey undercoverage. However, biases exist in the estimates to the extent that missed persons in missed households or missed persons in interviewed households have different characteristics from those of interviewed persons in the same age-sex-race-Hispanic group. Further, the independent population controls used have not been adjusted for undercoverage in the 1980 census.

For additional information on nonsampling error including the possible impact on CPS data when known, refer to Statistical Policy Working Paper 3, *An*

Error Profile: Employment as Measured by the Current Population Survey, Office of Federal Statistical Policy and Standards, U.S. Department of Commerce, 1978 and Technical Paper 40, *The Current Population Survey: Design and Methodology,* Bureau of the Census, U.S. Department of Commerce.

Sampling variability. The standard errors given in the following tables are primarily measures of sampling variability, that is, of the variations that occurred by chance because a sample rather than the entire population was surveyed. The sample estimate and its standard error enable one to construct a confidence interval, a range that would include the average results of all possible samples with a known probability. For example, if all possible samples were selected, each of these being surveyed under essentially the same general conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then approximately 90 percent of the intervals from 1.6 standard errors below the estimate to 1.6 standard errors above the estimate would include the average result of all possible samples.

The average estimate derived from all possible samples is or is not contained in any particular computed interval. However, for a particular sample, one can say with specified confidence that the average estimate derived from all possible samples is included in the confidence interval.

Some statements in the report may contain estimates followed immediately by a number in parentheses. For those statements one has only to add to and subtract from the estimate the number in parentheses to calculate upper and lower bounds of the 90 percent confidence interval. For example, if a statement contains the phrase "grew by 1.7 percent (±1.0)" the 90-percent confidence interval for the estimate, 1.7 percent, would be from 0.7 percent to 2.7 percent.

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The most common type of hypothesis appearing in this report is that the population parameters are different. An example of this would be comparing the number of children using computers to the number of adults using computers.

Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical. All statements of comparison in the text have passed a hypothesis test at the 0.10 level of significance or better. This means that, for most differences cited in the text, the absolute value of the estimated difference between characteristics is greater than 1.6 times the standard error of the difference.

Comparability of data. Data obtained from the CPS and other sources are not entirely comparable. This is due in large part to differences in interviewer training and experience and in differing survey processes. This is an additional component of error not reflected in the standard error tables. Therefore, caution should be used in comparing results between these different sources.

Note when using small estimates. Summary measures (such as medians and percent distributions) are shown only when the base is 75,000 or greater. Because of the large standard errors involved, there is little chance that summary measures would reveal useful information when computed on a smaller base. Estimated numbers are shown, however, even though the relative standard errors of these numbers are larger than those for corresponding percentages. These smaller estimates are provided primarily to permit such combinations of the categories as serve each data user's needs. Also, care must be taken in the interpretation of small differences. For instance, even a small amount of nonsampling error can cause a borderline difference to appear significant or not, thus distorting a seemingly valid hypothesis test.

Standard error tables and their use. In order to derive standard errors that would be applicable to a large number of estimates and could be prepared at a moderate cost, a number of approximations were required. Therefore, instead of providing an individual standard error for each estimate, generalized sets of standard errors are provided for various types of characteristics. As a result, the sets of standard errors provided give an indication of the order of magnitude of the standard error of an estimate rather than the precise standard error.

The figures presented in tables B-1 through B-4 are approximations to the standard errors of various estimates for persons, families and households shown in this report. To obtain the approximate standard error for a specific characteristic, the appropriate standard error in tables B-1 through B-4 must be multiplied by the factor for that characteristic given in table B-5. These factors must be applied to the generalized standard errors in order to adjust for the combined effect of the sample design and the estimating procedure on the value of the characteristic.

Standard errors for intermediate values not shown in the generalized tables of standard errors (B-1 through B-4) may be approximated by linear interpolation.

Two parameters (denoted "a" and "b") are used to calculate standard errors for each type of characteristic; they are presented in table B-5. These parameters were used to calculate the standard errors in tables B-1 through B-4 and to calculate the factors in table

Table B-1. Standard Errors of Estimated Numbers
(Numbers in thousands)

Size of estimate	Total, White, and non-Hispanic	Black and other	Hispanic
10	5	5	5
25	8	8	7
50	11	11	11
100	15	16	15
250	24	25	25
500	34	36	39
1,000	48	50	62
5,000	106	105	233
10,000	149	136	442
25,000	227	135	(X)
50,000	301	(X)	(X)
100,000	362	(X)	(X)
125,000	364	(X)	(X)
150,000	349	(X)	(X)

X Not applicable.
Note: For regional estimates, multiply the above standard errors by 0.94, 0.95, 0.94, and 0.90 for the Northeast, Midwest, South, and West, respectively. For a particular characteristic, see table B-5 for the appropriate factor to apply to the above standard errors.

B-5. They also may be used directly to calculate the standard errors for estimated numbers and percentages. Methods for computation are given in the following sections.

Standard errors of estimated numbers. The approximate standard error, S_x, of an estimated number shown in this report can be obtained in two ways. It may be obtained by use of the formula

$$S_x = fs \qquad (1)$$

where f is the appropriate factor from table B-5, and s is the standard error on the estimate obtained by interpolation from table B-1. Alternatively, the standard error may be approximated by formula (2) from which the standard errors in table B-1 were calculated. Use of this formula will provide more accurate results than the use of formula (1) above.

$$S_x = \sqrt{ax^2 + bx} \qquad (2)$$

Here x is the size of the estimate, and a and b are the parameters in table B-5 associated with the particular characteristic. When calculating standard errors for numbers from cross-tabulations involving different

Table B-2. Standard Errors of Estimated Percentages: Total, White, and Non-Hispanic

Base of estimated percentage (thousands)	Estimated percentage							
	1 or 99	2 or 98	5 or 95	10 or 90	20 or 80	25 or 75	35 or 65	50
10	4.8	6.7	10.5	14.4	19.2	20.8	22.9	24.0
25	3.0	4.3	6.3	9.1	12.2	13.2	14.5	15.2
50	2.1	3.0	4.7	6.4	8.6	9.3	10.3	10.8
100	1.5	2.1	3.3	4.6	6.1	6.6	7.2	7.6
250	1.0	1.4	2.1	2.9	3.8	4.2	4.6	4.8
500	0.7	1.0	1.5	2.0	2.7	2.9	3.2	3.4
1,000	0.5	0.7	1.0	1.4	1.9	2.1	2.3	2.4
5,000	0.2	0.3	0.5	0.6	0.9	0.9	1.0	1.1
10,000	0.2	0.2	0.3	0.5	0.6	0.7	0.7	0.8
25,000	0.10	0.13	0.2	0.3	0.4	0.4	0.5	0.5
50,000	0.07	0.10	0.2	0.2	0.3	0.3	0.3	0.3
100,000	0.05	0.07	0.10	0.14	0.2	0.2	0.2	0.2
125,000	0.04	0.06	0.09	0.13	0.2	0.2	0.2	0.2
150,000	0.04	0.05	0.09	0.12	0.2	0.2	0.2	0.2

Note: For a particular characteristic, see table B-5 for the appropriate factor to apply to the above standard errors.
For regional estimates, multiply the above standard errors by 0.94, 0.95, 0.94 and 0.90 for the Northeast, Midwest, South, and West, respectively.

Table B-3. Standard Errors of Estimated Percentages: Black and Other Races

Base of estimated percentage (thousands)	Estimated percentage							
	1 or 99	2 or 98	5 or 95	10 or 90	20 or 80	25 or 75	35 or 65	50
10	5.1	7.1	11.1	15.3	20.4	22.1	24.3	25.5
25	3.2	4.5	7.0	9.7	12.9	14.0	15.4	16.1
50	2.3	3.2	5.0	6.8	9.1	9.9	10.9	11.4
100	1.6	2.3	3.5	4.8	6.4	7.0	7.7	8.1
250	1.0	1.4	2.2	3.1	4.1	4.4	4.9	5.1
500	0.7	1.0	1.6	2.2	2.9	3.1	3.4	3.6
1,000	0.5	0.7	1.1	1.5	2.0	2.2	2.4	2.6
5,000	0.2	0.3	0.5	0.7	0.9	1.0	1.1	1.1
10,000	0.2	0.2	0.4	0.5	0.6	0.7	0.8	0.8
25,000	0.10	0.14	0.2	0.3	0.4	0.4	0.5	0.5

Note: For a particular characteristic, see table B-5 for the appropriate factor to apply to the above standard errors.
For regional estimates, multiply the above standard errors by 0.94, 0.95, 0.94 and 0.90 for the Northeast, Midwest, South, and West, respectively.

characteristics, use the factor or set of parameters for the characteristic which will give the largest standard error.

Illustration of the computation of the standard error of an estimated number. Text table A shows that there were 39,901,000 students in public school. Using formula (2), and the parameter, a = -0.000010 and b = 2,312 from table B-5, the estimate of the standard error is

$$S_x = \sqrt{(-0.000010)(39{,}901{,}000)^2 + (2{,}312)(39{,}901{,}000)} = 276{,}000^1$$

The 90-percent confidence interval for the number of students in public school is 39,459,400 to 40,342,600 (using 1.6 times the standard error). Therefore, a conclusion that the average estimate derived from all possible samples lies within a range computed in this way would be correct for roughly 90 percent of all possible samples.

Standard errors of estimated percentages. The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends upon both the size of the percentage and the size of the total upon which this percentage is based. Estimated percentages are relatively more reliable than the corresponding estimates of the numerators of the percentages, particularly if the percentages are 50 percent or more. When the numerator and denominator of the percentage are in different categories, use the factors or parameters from table B-5 indicated by the numerator. The approximate standard error,

[1]Using formula (1), the appropriate factor from table B-5 and a standard error obtained by interpolation from table B-1, the approximate standard error is (1.0) (271,000) = 271,000.

Table B-4. Standard Errors of Estimated Percentages: Hispanic

Base of estimated percentages (thousands)	Estimated percentage							
	1 or 99	2 or 98	5 or 95	10 or 90	20 or 80	25 or 75	35 or 65	50
10	6.2	8.7	13.6	18.7	24.9	27.0	29.7	31.1
25	3.9	5.5	8.6	11.8	15.7	17.0	18.8	19.7
50	2.8	3.9	6.1	8.4	11.1	12.0	13.3	13.9
100	2.0	2.8	4.3	5.9	7.9	8.5	9.4	9.8
250	1.2	1.7	2.7	3.7	5.0	5.4	5.9	6.2
500	0.9	1.2	1.9	2.6	3.5	3.8	4.2	4.4
1,000	0.6	0.9	1.4	1.9	2.5	2.7	3.0	3.1
5,000	0.3	0.4	0.6	0.8	1.1	1.2	1.3	1.4
10,000	0.2	0.3	0.4	0.6	0.8	0.8	0.9	1.0

Note: For a particular characteristic, see table B-5 for the appropriate factor to apply to the above standard errors.
For regional estimates multiply the above standard errors by 0.94, 0.95, 0.94 and 0.90 for the Northeast, Midwest, South, and West, respectively.

S(x,p), of an estimated percentage can be obtained by use of the formula:

$$S_{(x,p)} = fs \tag{3}$$

In this formula, f is the appropriate factor from table B-5 and s is the standard error on the estimate from tables B-2, B-3, or B-4. Alternatively, the standard error may be approximated by the following formula from which the standard errors in tables B-2, B-3 and B-4 were calculated. Use of this formula will give more accurate results than use of formula (3) above.

$$S_{(x,p)} = \sqrt{\frac{b}{x} p\ (100\text{-}p)} \tag{4}$$

Here x is the size of the subclass of persons or households which is the base of the percentage, p is the percentage (0 is the parameter in table B-5 associated with the particular characteristic in the numerator of the percentage.

Illustration of the computation of the standard error of a percentage Suppose that of the 8,085,000 full-time college students, 3,325,000 or 41.1 percent, use a computer anywhere. From table B-5, the appropriate "b" parameter is 2,312. Using formula (4), the approximate standard error of 41.1 percent is

$$S_{(x,p)} = \sqrt{(2{,}312/8{,}085{,}000)\ (41.1)\ (100\text{-}41.1)} = 0.8 \text{ percent.}^2$$

This means that the 90-percent confidence interval for the percentage of full-time students using a computer anywhere is from 39.8 to 42.4 percent, i.e., 41.1 ± (1.6 x 0.8).

Standard error of a difference. For a difference between two sample estimates, the standard error is approximately equal to

$$S_{(x\text{-}y)} = \sqrt{S_x^2 + S_y^2} \tag{5}$$

where Sx and Sy are the standard errors of the estimates x and y, respectively. The estimates can be of numbers, percentages, ratios, etc. This will represent the actual standard error quite accurately for the difference between two estimates of the same characteristic in two different areas, or for the difference between separate and uncorrelated characteristics in the same area. If, however, there is a high positive (negative) correlation between the two characteristics, the formula will overestimate (underestimate) the true standard error.

Illustration of the computation of the standard error of a difference. Suppose that there were 683,000 full-time college students using computers at home and that there were 541,000 part-time college students using computers at home. The apparent difference is 142,000. Using formula (2) and the appropriate parameters from table B-5, the approximate standard errors of these two estimates are 40,000 and 35,000, respectively.[3] Therefore, from formula (5), the approximate standard error of the estimated difference of 142,000 persons is

$$S_{(x\text{-}y)} = \sqrt{(40{,}000)^2 + (35{,}000)^2} = 53{,}000.$$

This means that the 90-percent confidence interval for the true difference between full-time college students using computers at home and part-time college students using computers at home is from 57,200 to 226,800. Therefore, a conclusion that the average estimate of the difference derived from all possible samples lies within a range computed in this way would be correct for roughly 90 percent of all possible samples. Since this interval does not contain zero, we can conclude with 90 percent confidence that the number of full-time college students using a computer at home is greater than the number of part-time college students using a computer at home.

[2]Using formula (3), the appropriate factor from table B-5 (1.0), and a standard error from table B-1, B-2 or B-3, the approximate standard error is (1.0) (1.1) = 1.1 percent.

[3] $\sqrt{(-0.000010)\ (683{,}000)^2 + (2{,}312)\ (683{,}000)} = 40{,}000$, and $\sqrt{(-0.000010)\ (541{,}000)^2 + (2{,}312)\ (541{,}000)} = 35{,}000$.

of the difference derived from all possible samples lies within a range computed in this way would be correct for roughly 90 percent of all possible samples. Since this interval does not contain zero, we can conclude with 90 percent confidence that the number of full-time college students using a computer at home is greater than the number of part-time college students using a computer at home.

Table B-5. Standard Error Parameters and Factors

Characteristic	Parameter a	Parameter b	f factor
Persons			
Total, White, and non-Hispanic:			
Enrolled in school	-0.000010	2,312	1.0
Household type, age of householder, presenceof children	-0.000115	4,480	1.4
Unemployed	-0.000015	2,206	1.0
Black and other:			
Enrolled in school	-0.000075	2,600	1.0
Household type, age of householder, presence of children	-0.000186	6,426	1.6
Unemployed	-0.000073	2,536	1.0
Hispanic:			
Enrolled in school:			
Levels	+0.001744	2,131	1.0
Percentages	(X)	3,873	1.0
Household type, age of householder, presence of children			
Levels	-0.000330	5,673	1.6
Percentages	(X)	11,414	1.7
Unemployed	-0.000108	2,087	1.0
Families			
Total, White, and non-Hispanic:			
Household type, age of householder, presence of children	-0.000010	1,778	0.9
Household income	-0.000010	1,896	0.9
Employment status and occupation of householder	-0.000025	2,013	0.9
Unemployed	-0.000015	2,206	1.0
Black and other:			
Household type, age of householder, presence of children	-0.000046	1,606	0.8
Household income	-0.000060	2,067	0.9
Employment status and occupation of householder	-0.000058	2,013	0.9
Unemployed	-0.000073	2,536	1.0
Hispanic origin:			
Household type, age of householder, presence of children	-0.000106	1,820	0.9
Household income	-0.000120	2,067	1.0
Employment status and occupation of householder	-0.000108	1,863	0.9
Unemployed	-0.000108	2,087	1.0

X Not applicable.

Note: For regional estimates multiply the "a" and "b" parameters by 0.88, 0.91, 0.89 and 0.81 for the Northeast, Midwest, South, and West, respectively.